Research on Modern Higher Education 2, 02014 (2017)
DOI: 10.24104/rmhe/2017.02.02014

The distribution of scientific index weight based on information entropy and correlation coefficient

Rui Chi*, Xiang Su & Nianxin Wang
School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, China

ABSTRACT: The weights of scientific research indexes affect the accuracy of assessment. Information entropy reflects the amount of index information; correlation coefficients reflect the correlation between the indexes and assessment results. In this paper, it takes the normalized products of information entropy and correlation coefficients as the weights of scientific assessment indexes. The result of empirical analysis shows the reason why "The Core Journals" (native) and foreign journals being assigned with the largest weight coefficient is because both of its entropy and correlation coefficients are big. But some important index weights of the national projects and "A Journal Papers" are smaller because only a few people have the index data, and this will result in small index information entropy. It's last verified by the sample distribution that the product of the scientific index and the weight coefficient is conducive to grading the assessment of the teachers.

Keywords: scientific assessment; weight coefficient; amount of index information; sample distribution; index system; assessment result

1 INTRODUCTION

Performance reform is an important assignment in the current reform of colleges and universities. Research assessment plays an important role in the performance reform. Fair and accurate research evaluation determines the success of the performance reform. Research work is usually evaluated with a number of indexes. Current index options mostly refer to the relevant departments of the Ministry of Education's subject review, project application and so on. Different evaluation indexes play different roles, so the weights assigned to them are different.

The existing distribution methods of the weight of indexes can be divided into three categories according to its attributes. The first one is a method of subjective weight coefficient containing the AHP, Delphi method and expert judgment method. The second one is the objective weight coefficient method, such as fuzzy clustering analysis, entropy coefficient, the main component analysis and maximizing deviations method. The last one is a composite method of subjective and objective weight coefficient. In order to more accurately determine a reasonable weight, the domestic scholars have been committed to the exploration of various methods to determine the weights [1-3].

Information entropy of index reflects the information amount of the index. The correlation coefficient between index and assessment result reflects their relevance. Some indexes may be sufficient in information but not related to the assessment result, or they may be correlate, but with little information. Thus the indexes will not play a big role in the assessment. So greater weight will be distributed to assessment index which has large information entropy and high correlation coefficients in the assessment result. According to the theory of information entropy and correlation coefficient, redistribution weight method for research index was explored with scientific assessment index in a university.

2 THE CONCEPT OF INFORMATION ENTROPY AND CORRELATION COEFFICIENT

2.1 *The concept of information entropy*

The concept of entropy is proposed for the first time in

*Corresponding author: chirui1978@163.com

1856 by the physicist R. Clausius and used to quantitatively explain the second law of thermodynamics. In 1948, CE. Shannon employed it to measure the uncertainty of information entropy in the information theory. Assuming that there is a discrete and random event X, the probability of discrete event state x_i is p_i. Then $log_a \frac{1}{p_i}$ is the amount of information in the x_i state[4] . The information amount contained in different states is different, so the information amount is a random variable (related to its probability). But the uncertainty of whole event *X* cannot be measured with it. So the mathematical expectation of self-information in discrete state is defined as the average amount of information $H_r(X)$ which is known as the information entropy, i.e.

$$H_r(X) = -\sum_{i\geq 1} p_i log_r \frac{1}{p_i} \tag{1}$$

The information entropy generally reflects the average information of discrete random event *X* provided by each state x_i[5]. The information entropy has the following properties.

(1) Non-negativity. $H(X) \geq 0$, the condition for equality is that if and only p_i =1 for some *i*;

(2) Additivity. For independent events, and the sum of entropy is equal to the entropy of sum;

(3) Extremum. When the state is in equal probability, that is p_i=1/*n*, the entropy is maximum $H(X) = -\sum_{i\geq 1} p_i log_a \frac{1}{p_i}$ =$log_a n$;

(4) Symmetry. The sequence of variable $(x_1, x_2, x_3 ...)$ exchanges randomly, but the entropy of random events X is constant value.

According to the characteristics of the information entropy, the more scientific research indexes are (the quantity of state x_i is large), the more uniform probability and the bigger information entropy could be; on the contrary, the information entropy is smaller [6]. So the scientific assessment should be based on the indexes that the majority of teachers have. The scientific research level can be distinguished according to the difference of these indexes.

2.2 *The concept of correlation coefficient*

For random variables *X* and *Y*, variance of *X* is $E\{[X-E(X)]^2\}$, and variance of Y is $E\{[Y-E(Y)]^2\}$. The covariance of *X* and *Y* is *Cov(X, Y)* = *E{[X-E(X)][Y-E(Y)]*, and then the correlation coefficient ρ_{xy} of random variables *X* and *Y* is defined as:

$$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{2}$$

Correlation coefficient ρ_{xy} indicates the close degree of the linear relationship between *X* and *Y*. When $|\rho_{xy}|$ is big, *X* is highly related to *Y*. When $|\rho_{xy}|$ is small, *X* and *Y* are poorly correlated. When $\rho_{xy} = 1$, *X* and *Y* are in linear relationship with the probability 1. When the $\rho_{xy} = 0$, *X* and *Y* are not related. When $\rho_{xy} > 0$, *X* and *Y* are in positive correlation. When $\rho_{xy} < 0$, X and *Y* are in negative correlation [7].

3 WEIGHT ALGORITHM DESIGN OF SCIENTIFIC RESEARCH ASSESSMENT INDEX

Research assessment is a multi-attribute decision making problem. The incommensurability of the indexes has important effect on the assessment results [8]. The impact of assessment index on the assessment results depends on two aspects. The first is that if the assessment indexes have sufficient amount of information, and the second is that if the assessment indexes are related with the assessment results.

The information amount of assessment indexes can be measured with the information entropy and the correlation between the assessment indexes and the assessment results can be measured with the correlation coefficient. Therefore, this paper proposes a method using the combination of information entropy and correlation coefficient to determine the weights of assessment indexes. The specific steps are as follows:

(1) According the distribution range of the k_th assessment index X_k, D intervals are divided. If the index value is an integer value, it should take the maximum integer; if the maximum value is less than 10, D is set as the integer value. If the maximum integer is greater than 10, D is set as 10; if the index value is a real number, D is also set as 10.

(2) Set the number of samples falling into the i_th interval is n_i and the total sample number is *N*. The probability of the samples falling into the interval is p_r=$\frac{n_i}{N}$. The index information entropy $H_r(X_k)$ can be obtained from formula (1).

(3) Calculate the correlation coefficient $\rho_{x_k y}$ between the K_th assessment index X_k and assessment result *Y*.

(4) Multiply the information entropy of K_th assessment index $H_r(X_k)$ by correlation coefficient $\rho_{\text{x}_\text{k}\text{y}}$, which is $U_k = H_r(X_k)*\rho_{x_k y}$.

(5) Obtain the normalized weight coefficient of the assessment index according to the U value of all indexes.

$$W_K = \frac{U_k}{\sum U} \tag{3}$$

4 DESIGN OF EMPIRICAL RESEARCH

To meet the needs of scientific research performance assessment in a certain university, a case study is carried out with the scientific research data in last several years.

4.1 *Selection of the assessment index*

Scientific research achievements can be divided into five categories: research projects, research papers, works, research awards, patents. These five categories can be taken as first level of the index. The index of patents is deleted as the quantity of this kind of index from School of Economics and Management is very small. The selection of the second level of the index is based on the statistical index of the Ministry of Education Science and the characteristics of Management Science [9], see Table 1.

Table 1. Assessment index system of teachers' scientific research performance

First level of the index	Second level of the index
Research projects	Number of national projects
	Number of provincial and ministerial level projects
	The city level project funds
	Cooperation projects
Research papers	Papers indexed by SSCI/SCI/SCIE
	A Journal by Management Science Department of the National Natural Science Foundation
	B Journal by Management Science Department of the National Natural Science Foundation
	Paper indexed by CSSCI
	Core / Foreign Periodicals
	General journal
Works	works
	Edited books
Research awards	awards*

* The specification of different scores of the research awards is as follows: (1) the first prize on provincial-level incentives, 3 points; (2) the second prize on provincial-level B incentives, 2 points; (3) the third prize on provincial-level incentives, 1 point; (4) the first prize on city departmental-level incentives, 0.8 point; (5) the second prize on city departmental- level incentives, 0.6 point; (6) the third prize on city departmental-level incentives, 0.4 point.

Other science indexes and the data of assessment are shown in Table 2.

Based on the methods of expert assessment described above, an appraisal report of 59 full-time teachers was made. The results were in four grades: excellent, good, middle, and pass, but no fail grade. The four assessment results were corresponded to the 5, 4, 3, 2 four assessment levels.

4.2 *Results and analysis*

The information entropy of the 13 indexes is shown in Figure 1.

It can be seen from Figure 1 that there are large differences of information entropy of different indexes. The highest information entropy is the native Core Journals or foreign journals, which is followed by common journals and CSSCI journals. The reason of the big information entropy of these indexes is because the distribution of the indexes is relatively dispersed. After dividing the intervals, the number of samples in the interval is relatively uniform, in other words, the index values of different people are different, and this provides a large amount of information. The index with minimum information entropy is edited book as only a few people have edited books, and the amount of information in the assessment is limited.

The correlation coefficient of the 13 assessment indexes and assessment results is shown in Figure 2.

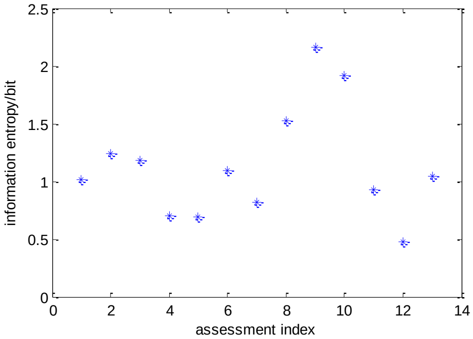

Figure 1. Assessment indexes of information entropy

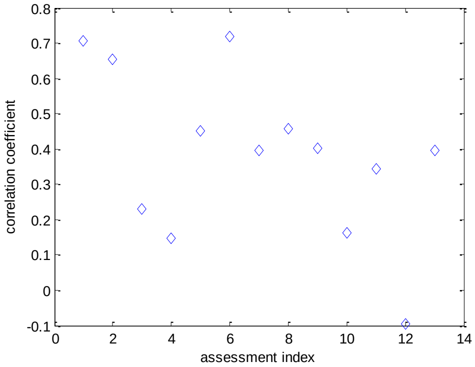

Figure 2. The correlation coefficients of the assessment indexes and the results

It can be seen from Figure 2 that the number of published papers (A Journal), national projects and provincial projects and the diploma are the highest indexes in the correlation coefficient of assessment index and the assessment results. In addition, the correlation coefficient of edited books and the scientific assessment is negative, that is they are in negative correlation. The negative correlation lies in no impact of edited books on the scientific assessment, and as a few samples with lower scores have edited books, the correlation between the edited books and assessment

Table 2. A group of assessment indexes and corresponding assessment results

No.	A	B	C	D	E	F	G	H	I	J	K	L	M	AL*
1	0	1	0.5	7	0	0	0	3	1	1	2	0	0.4	4
2	0	1	0.5	2.5	0	0	0	0	14	0	0	1	0.4	4
3	0	0	0	0	0	0	0	0	0	1	0	0	0	2
4	0	0	1.1	39	0	0	0	1	1	3	0	0	0.4	2
5	1	2	0	2	0	3	0	1	11	3	1	0	2.8	5
6	0	0	0	21.5	0	0	0	0	5	7	0	1	0	3
7	0	0	0	0	0	0	0	0	3	0	0	0	0	2
8	0	0	2.2	0.4	0	0	0	2	2	2	0	0	0	3
9	0	0	2.5	0.4	0	0	0	0	0	1	0	0	0	2
10	1	0	1	0	0	0	0	0	0	0	0	0	0	4
11	0	0	0	2	0	0	0	0	0	0	0	0	0	2
12	0	1	0.6	0	0	0	0	0	0	1	1	0	0	3
13	0	0	1.5	0.8	0	0	0	0	1	5	1	1	0	3
14	0	0	0	0.4	0	0	0	0	0	2	0	0	0	2
15		1	0	0	0	0	0	0	0	0	0	0	0	3
16	0	0	0.5	0	0	0	0	0	1	0	1	0	0	3
17	0	1	2.8	0	0	0	0	0	1	0	1	0	0	4
18	0	0	0	0.4	0	0	0	0	3	3	0	0	0	3
19	0	0	0	0	2	0	0	0	1	1	0	0	0	3
20	0	0	0.8	0	0	0	0	0	0	0	0	0	0	2
21	0	0	0.25	0.4	0	0	0	1	4	3	0	0	0.7	3
22	0	0	0.4	6.4	0	0	0	0	0	0	0	0	0	3
23	0	0	0.85	0	0	0	0	0	1	2	0	0	0	3
24	0	1	0	0	0	0	0	6	2	14	2	0	0	4
25	0	0	0.5	4	0	0	0	0	4	8	0	1	0	3
26	1		1.3	0.8	0	0	0	0	0	0	0	0	0	4
27		1	1.7	1.4	0	4	2	3	8	17	1	0	0.6	5
28	1	2	2.05	0.4	0	2	1	3	6	5	0	0	0	5
29	1	1	0	0.4	1	0	0	0	0	1	0	0	0	5
30	1	3	0.6	2.9	2	0	1	1	5	0	0	0	0	5

* A, Number of national projects; B, Number of provincial and ministerial level projects; C, city level project funds; D, Cooperation projects; E, Paper indexed by SSCI/SCI/SCIE; F, A Journal by management science department of the National Natural Science Foundation; G, B Journal by management science department of the National Natural Science Foundation; H, Paper indexed by CSSCI;I, Core / Foreign Periodicals; J, General journal; K, works; L, Edited book; M, awards; AL, assessment level.

results is negative finally. In this case, the correlation coefficient of these samples should be set as 0. The weight coefficient is shown in Figure 3.

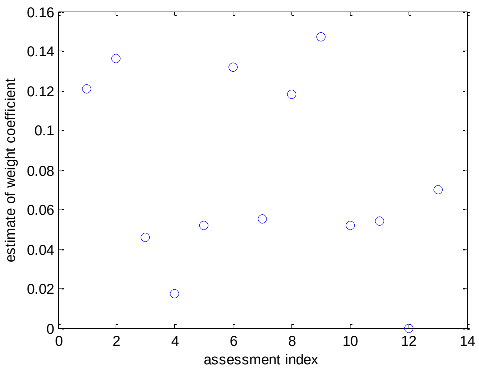

Figure 3. Estimation of the weights of assessment indexes

The six biggest indexes of the weight coefficient are the native Core Journals or the foreign journals, provincial-level projects, A journals, national projects, CSSCI indexed journals, and the diploma. The difference of information entropy of different indexes is relatively small, while the difference of correlation coefficients is relatively large. So the weight coefficient mainly lies in the relevance and it should be combined with the information entropy to distribute the weights of the indexes. If it only considers the relevance, the information entropy is small, and the differences between the assessment samples cannot be distinguished. Likewise, if the information entropy is large but the correlation is small, there is no significance to the assessment result.

Also it can be seen that the weight coefficient is not equal to the proportion of the index. For example, the proportion of national projects and A Journals in all indexes is higher than that of the native Core Journals or foreign journals; however, the weight coefficient of national projects and A journals is less than that of the native Core Journals or foreign journals. It is because only a few people obtain the scores of national projects and A Journals and the amount of information of these indexes is small. These indexes are not conducive to grading the assessment results. Conversely, almost every sample can obtain the scores of the native Core Journals or foreign journals and there are differences in the indexes value of teachers, so that the indexes are conducive to grading the assessment results.

The pros and cons of weight coefficient can be measured by the distribution of the index multiplied by the weight coefficient. It should first normalize each index of the sample to [0, 1], and then multiply the index of the sample by the weight coefficient, by then, the distribution of the samples can be estimated accordingly. The sample distribution of weight coefficient performed by the above methods is as shown in Figure 4.

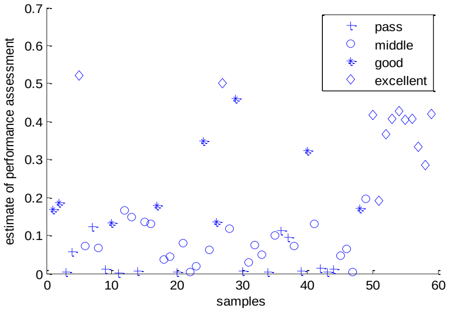

Figure 4. The distribution of samples after the indexes being weighted

The assessment levels in Figure 4 were based on year-end awards and the voting results from the experts in the last year. The scores on longitudinal coordinates are the products of the performance indexes of each sample and the corresponding weight coefficients, and it can be seen that the sample distribution matches well with the assessment results, indicating that the weight coefficient is reasonable. The correlations between the selected indexes are small, and the redundancy of the index information is also small, so the weighted assessment indexes can be employed in the scientific research assessment with SVM, and so on. The accuracy of the test result was greatly improved.

5 CONCLUSIONS

(1) The information entropy of scientific research index reflects the amount of information for scientific research assessment, and the correlation coefficient of related assessment result of scientific research index reflects the correlation between the index and the assessment results. The combination of these two factors can estimate the weight distribution coefficient of the index.

(2) The product of the information entropy of the index and the correlation coefficient of the assessment result can be used as the coefficient of the assessment result. The sample distribution after the index being weighted can be applied to measure the advantages and disadvantages of the weight coefficient. In this paper, the weight coefficient makes the sample distribution of the weighted index match will with the assessment result, and it is conducive to the training of vector machine and improving the test accuracy.

(3)According to the empirical testification, we found that the indexes with large weight coefficients are always the indexes most teachers have, and the indexes of different teachers are significantly different. On the contrary, the weight coefficients of the indexes that only a few teachers have are relatively small.

(4)The indexes with large weight coefficients do not indicate they are the most important ones. Large weight coefficients are conducive to differentiating the grades evaluation. But the most important index can only contribute to excellent assessment result.

ACKNOWLEDGEMENTS

The paper is sponsored by Project of the National Natural Science Foundation of China (71471079), Zhenjiang Soft Science Research Project (RK2014036) and Research and Practice on the Teaching Reform of Graduate Education in Jiangsu University of Science and Technology (YJG2014Y_09)

REFERENCES:

[1] Jin Jia-jia, MI Chuan-min, Xu Wei-xuan, WANG Qun-feng & Wei Hengwu. 2012. The maximum entropy empowerment model for evaluating index considering the expert evaluation information. *Chinese Journal of Management Science*, 20(2): 135-143.

[2] Zhang Hone, Zhao Zhen & Zhao Xing-cun. 2012. The comprehensive factors and judge weights of the factors of self-evaluation system for research assessment exercises in the medical colleges and universities in China. *Science and Technology Management Research*, (16): 102-105.

[3] Wei Li, Guofeng Chen & Cheng Duan. 2010. Research and implementation of index weight calculation model for power grid investment returns. *Lecture Notes in Computer Science*, 6138: 44-52.

[4] Yao Xiaoyang, Sun Xiaoleis, Wu Dengsheng & Yang Yuying. 2015. Correlation research of country risk based on mutual information. *Systems Engineering-Theory & Practice*, 35(7): 1657-1665.

[5] Zhang Chunqin, Juan Zhicai & Jing Peng. 2015. Evaluating the efficiency of urban public transit operators using information entropy and SE-DEA combined model. *Industrial Engineering and Management*, 20(1): 146-153.

[6] Zhang Xinlei. 2015. Research on Information entropy methods in risk measure. Beijing: Beijing Jiaotong University.

[7] Dong Yuehua & Liu Li. 2015. An optimized algorithm of decision tree based on correlation coefficients. *Computer Engineering & Science*, 37(9): 1783-1793.

[8] Li Qinyang, Wang Jinyan & Jiao weihong. 2015. Evaluation of a provincial competitiveness information entropy based on interval number TOPSIS. *Statistics & Decision*, 426(6): 63-65.

[9] Qi Yong, Li Qianmu & Sun Haihua. 2009. The university innovation ability appraisal based on PCA -BP and cluster method. *Science of Science and Management of S.&T,* (10): 112-117.

[10] Wei Chen & Xiaohong Hao. An optimal combination weights method considering both subjective and objective weight Information in power quality evaluation. *Lecture Notes in Electrical Engineering*, (87): 97-105.